# Implementation Plan for the Unified Production Environment

**(DRAFT)**

**NERSC**
**1994**

1/18/95

# Implementation Plan for the Unified Production Environment

## Contents

# DEVELOPMENT, COMPUTING AND ASSIMILIATION

## Scope of Responsibility

*The Center is chartered to provide services so that the science pursued by the research community is not constrained by the lack of access either to capability or sophistication in the computational resources (hardware and software).* The Development, Computing, and Assimilation (or DCA) environment lies at the heart of the service to be provided by NERSC in its role as an access center for Energy Research scientists. NERSC will seek to provide a balanced and unified environment to accommodate the three phases of computational studies: the development of the application, the execution of the application on a production supercomputer, and the assimilation of the generated results.

## Area of Emphasis

A small number of diverse capability machines presents an optimal environment to serve all the high-end applications of the ER community. At the center of the Unified Production Environment (UPE) will be the capability systems, including the highest high-end machine (an MPP) plus additional high capability systems which complement the MPP (such as the C90 or other large memory systems). Surrounding this core must be an augmented supercomputing auxiliary service (SAS) which will provide the necessary services to allow the high-end to fully realize its potential. Part of this SAS environment must include an enhanced assimilation environment well-integrated into the UPE, tracking the leading edge both in capability of equipment and in techniques employed. Another part of the SAS environment should be a software rich development engine for codes which are targeted for production on the MPP. A very good candidate for this is a symmetric multiprocessor (SMP).

The following sections describe the three subsections of the DCA triad: Development Environment, Computing Environment and Assimilation Environment.

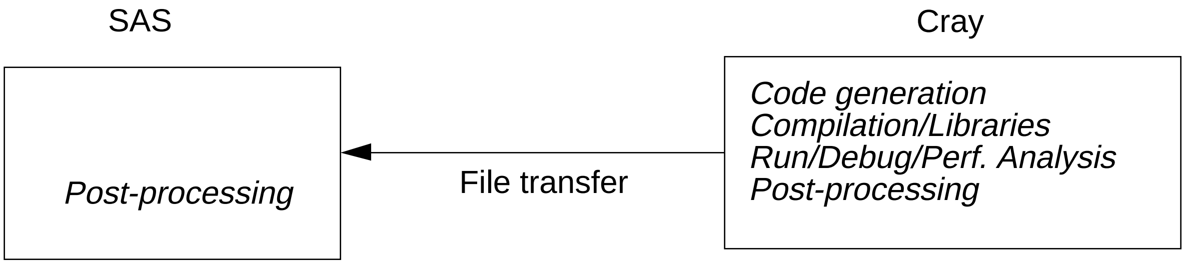
# Development Environment

***NERSC must build a development environment which extends across a collection of specialized resources, including SAS, the C90 and MPPs.***

The development environment consists of the interfaces and tools with which the scientist writes, debugs, and analyzes the performance of an application code. Lower end computational devices tend to have more efficient and mature development environments than higher end resources. There is consequently a natural tendency to develop at the lower end and run at the higher. Distributed computing software will be used to hide from the customer the fact that the services rendered, including software development, are being provided on multiple machines from different vendors.

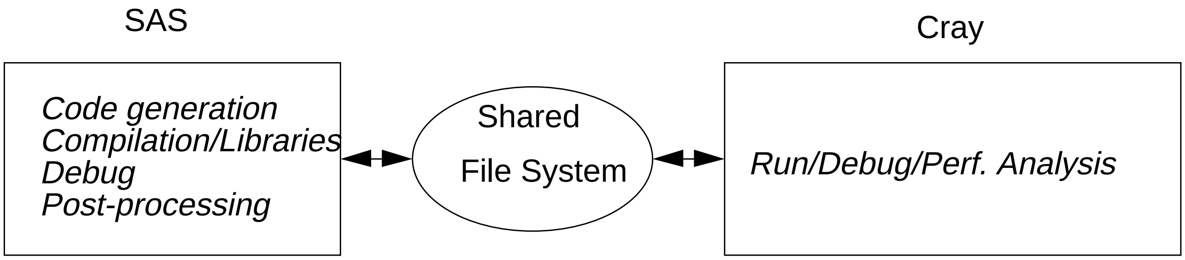
The following diagrams depict the transition from the development paradigms of yesterday and today to those we intend to realize in the next 3 years.

*Phase 0: Historical NERSC development paradigm*



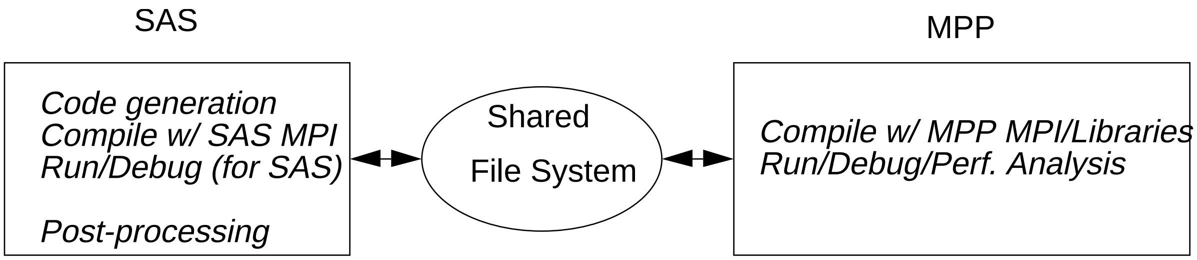
Characteristics: Little file sharing between systems; development done on Cray; some post-processing done on SAS. Slow response and limited tool set.

*Phase 1: Provide Cray C90 cross-development tools on SAS (by mid-1995)*



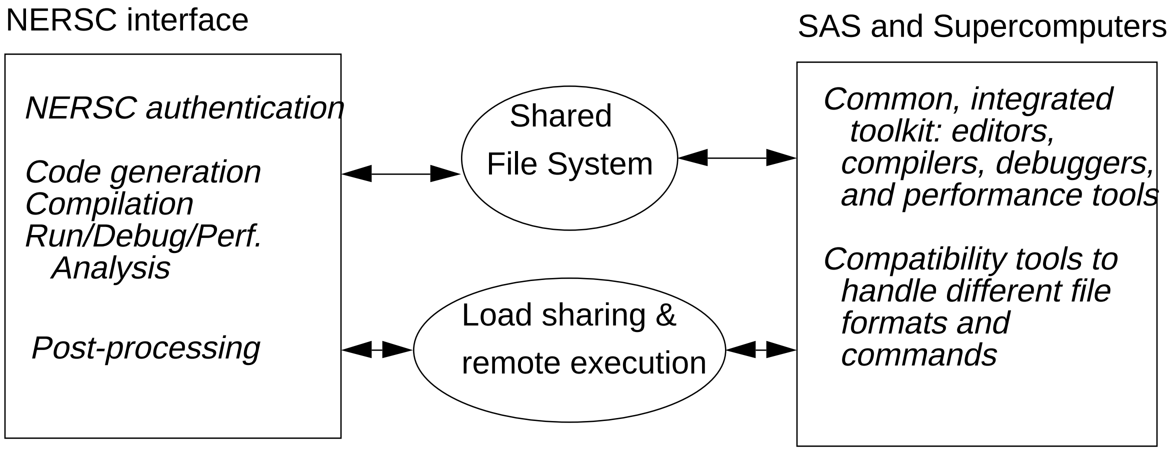
Characteristics: All NERSC customers have SAS accounts and can share files between Crays and SAS. Cray cross-development tools (compilers, library, debugger) provided on SAS creating compatible and more responsive development environment.

*Phase 2: Provide MPP development environments on MPP and SAS (by early 1996)*



<u>Characteristics</u>: Initial MPP development tools from PTools Consortium, commercial and public domain will lack sophistication of integrated toolkits implemented in Phase 1. As cross- and compatible compilers become available, they will be installed on SAS.

*Phase 3: Integrate development environment across SAS and supercomputers (by mid-1996)*



<u>Characteristics</u>: A single NERSC login; an integrated, common software development tool kit which extends across machine boundaries; remote execution of user and system processes to balance load across eligible platforms.

***An SMP should be provided as a major element of the DCA environment.***

An SMP with its rich software environment can be used as a multi-user DCA engine for debugging, postprocessing and code development for the MPP. In this environment, many users can coexist on the SMP, developing parallel applications which when mature will execute on the MPP. *It is extremely important that the programming model and the compilers be compatible across the SMP and MPP platforms.*

The SMP, therefore, sits at the interface between the high-end and the low-end: part supercomputer, part supercomputing auxiliary service (SAS), but providing the precise support to the high-end which is required to compensate for the current weaknesses of the MPP. *The SMP represents a major contributor to the stability and functionality of the UPE during the transition period, 1996-1998.*

***The Center will encourage the transition to parallel computing before the arrival of the PEP system.***

Since the programming environment on an MPP is very different from that on a standard vector supercomputer, unless the community develops some parallel applications and some skills now, the new machine will not be populated for some time with suitable applications. This can be achieved by providing access to parallel computers at the HPCRCs and at LLNL. This must be accompanied by appropriate consulting and collaborative services.

**Milestones**

*for Phase 0 (initiated by 1994 or early 1995)*

1. <u>Porting</u>: preceding the arrival of the PEP system (a system which will not have full production status) there must exist a project for the transition of capability applications from sequential and parallel systems to massively parallel systems. This should free up cycles during 1995-96 on the sequential-vector platforms. This work has been initiated. The Center is also isolating the most time intensive codes and working with users to effect parallel versions.

2. <u>Access to parallel computers before the arrival of the PEP</u>: This project is underway, and involves securing substantial time from two HPCRCs as well as from the LLNL T3D. An applications process is involved to gain access, and NERSC provides both consultation and collaboration in development of parallel applications.

3. <u>Special Parallel Processing (SPP) for the C90</u>: the balance between SPP dedicated time and interactive time will represent a balance between the requirements for capability and the needs of DCA users. This ratio may change as the MPPs off-load capability from the C90. The workload must be carefully monitored on both systems (the C90 and the MPP) in order to make the optimum choice for the magnitude of the SPP allocation.

*for Phase 1 (complete by mid-1995)*

1. <u>C90 for capability computing</u>: Off-load inappropriate work from C90 creating more cycles for capability. Off-load from C90 disks small files, such as active source trees, creating more space for capability

2. <u>Provide common home directories</u>: This should be effected across Crays and SAS for all NERSC users (see chapter, **Administration**)

3. <u>Enhance SAS development environment:</u> Install Cray software development tools, such as Craysoft, on SAS.

4. Documentation: Document desired development paradigm as part of “How to Use the UPE”

5. Partner with users: Work with selected users to refine environment and develop training in use of shared file system and development tools

*for Phase 2 (complete by early 1996)*

1. Common programming models across SAS and MPPs: Provide standard message passing software and high performance FORTRAN on SAS and MPP

2. Acquire an SMP: Install and integrate an SMP into the NERSC DCA environment as a part of SAS

3. Apply pressure on MPP vendor: Work with MPP vendors to encourage development of integrated toolkits

4. Partner with users: Enlist help of subset of users to test out development tools and develop training classes.

*for Phase 3 (complete by late 1996)*

1. Automated load sharing: Extend load sharing capability to include supercomputers for load sharing and remote execution.

2. POSIX compliance: Install POSIX-compliant UNIX and integrated tool kit across supercomputers and SAS:
   Compilers (F90, C, C++, HPF)
   Debugger (interactive, X-based)
   Libraries (CraylibSci, class libraries)
   Source code management tools

# Computing Environment

The Computing Environment in this document contains the following elements:

1. the primary ***hardware*** (including disk) placed on the NERSC floor for purposes of high-end computing;
2. the ***software*** residing on the computational platforms, including system software (such as batch and interactive workload schedulers that facilitate a successful multi-user execution environment) and the applications and library software (including engineering and chemistry codes as well as mathematical libraries).

We will address each of these two areas:

## (1) Hardware

***For the next few years the users will be best served by a capability production environment with a small number of architectures at the high end.***
Since the mission of the Center focuses on capability, the *most advanced and capable* resources must be made available for the research community.

- *The highest high-end computing service should be on an MPP.*
  NERSC must offer this service or forgo its charter to provide the most capable hardware to the user community. The MPP high-end will involve a phased procurement. In 1995, a smaller Pilot Early Production (PEP) machine will be put on the floor. This machine will consist of at least 128PEs (probably more) but will almost certainly feature an unsophisticated DCA environment. If the vendor succeeds in meeting a series of production status requirement (PSR) milestones designed to guaranteed an acceptable (but not stand-alone) DCA environment, the machine will be upgraded to at least 512 processors in about a year. This will represent the highest high-end at NERSC until 1998/1999.

- *The C90 will continue to carry a heavy capability workload.*
  The vast majority of NERSC capability codes currently run on the C90, many utilizing the Special Parallel Processing allocation. For problems which fit on this machine and for which a shared memory programming model is most appropriate, this machine can remain as a stable capability platform for several years.

- *An SMP should be considered as an element of the capability core.*
  Making reference to *Table 1* and *Figure 1* of the **Strategic Plan**, a single SMP will not compete with the MPP at the highest high-end, but represents a capability solution for those codes which struggle to map efficiently to a distributed memory architecture. The most troublesome codes for the MPP remain those employing both unstructured grids and implicit differencing and have to resort to mathematical techniques such as conjugant gradient. Some of these, will map more easily to the SMP or the C90, and *many would benefit from the large memory available on the SMP and which is no*t *available on the C90*. Many engineering application codes coming from independent

software vendors require a high capability engine, and *will be written for the SMP before versions are available for MPP.* We see the primary benefit of the SMP, however, as an element of SAS as is emphasized in the *Development* subsection (preceding).

To reach these goals we need to meet the following milestones concerning the mix of capability hardware on the NERSC floor in the coming three years.

**Hardware Milestones**

1. Architecture Mix: (By mid-1995) Acquire better projections from our customers of the NERSC capability workload requirements to determine the mix of capability machines and their number.

2. Cray 2s: (By mid-1995) Determine the role (if any) of the Cray-2s in supporting the high-end after 1995. This is related to the analysis to be undertaken in (1) above.

3. High-end SMP: (By mid-1995) NERSC should arrange with at least one SMP vendor to evaluate for a period of time (not to exceed one year) a fully-configured, high-end symmetric multiprocessor. The evaluation would be based in part on the work performed by a subset of users whose codes are better suited to a shared memory machine, and would include an assessment of the SMP as a development, computing and assimilation system for the high-end MPP.

4. MPP Configuration and Upgrade: (By mid-1996) The FCM will be the primary capability system at NERSC until 1998-1999. Hence, the FCM must have a balanced configuration of peripherals (disk, I/O nodes) and upgraded to track technological evolution. In the year after the arrival of the PEP, a small task force of NERSC staff and users will determine a balanced configuration for the FCM by analyzing user workloads on the PEP, and also recommend mid-life upgrade paths for the FCM technology.

## (2) Software

***The Center will endeavor to provide a complete DCA environment on each capability system as part of the effort to create a unified production environment (UPE) across all systems.***

Throughout its existence this Center has taken advantage of the following synergy: a capability system, wedded to a responsive and sophisticated software understructure, can also be utilized as a (DCA) system. This combination of compute capability with DCA capability provides both generality and flexibility in the range of production services provided by NERSC. Conversely, the Center has seen that *the high-end system that permits a DCA environment also expedites full use of capability*. Much of what has slowed the success of MPP to date is that the existing MPP environment has not permitted a DCA environment. This has made it difficult to run numerous small jobs in debug mode, to run a job of a certain size and processor count whenever the researcher wishes to do calibration calculations, or to do all of the interactive code manipulation chores needed to ready a code for a production run.

***To effectively utilize the capability platforms NERSC customers must have access to the appropriate engineering, chemistry and science applications, as well as a complete mathematical software library.***

Not all applications must live on all systems, but a complete mathematical software environment (including linear algebra) is a very important element of a production environment. The Center must make a major effort to assure that each capability system has a functional and reasonably complete mathematical software library.

Judgments concerning which (engineering and science) applications will be ported to (or acquired for) which capability platforms will be made on the basis of the cost of the software and its computational requirements. This could require a significant investment in staff effort and time.

**Software Milestones**

1. <u>MPP schedulers</u>: (By early 1996) Adequate space and time sharing schedulers must be developed by vendors with help, if necessary, from NERSC. The production environment on the fully configured system (FCM) must allow for two simultaneous modes of operation: an interactive mode for development, short runs and low processor count simulations and a batch mode directed at long running capability calculations.

2. <u>MPP operating environment</u>: (By early 1996) The Center should determine if the “Livermore model” of partitions and gang scheduling on the MPP will provide the needed interactive (DCA) environment. This decision should come from a small task force of NERSC staff and users which forms during the first year of use of the PEP machine. Task force members should recommend a style of environment for the MPP and from this propose a configuration (software, disk, I/O, etc.) for the FCM.

3. <u>Central User Bank</u>: (By late 1995) We must extend CUB to provide central user account and allocation management on the new machines, such as PEP, FCM and SMP. This will force a re-examination of the concept of a single currency (currently the *cru*) given the disparity between the MPP and the other capability platforms. We must implement policies within the context of CUB to ensure that a large central allocation intended primarily for use of one resource (MPP, for example) is not instead used to overtax another resource (C90, for example).

4. <u>Porting applications not supported by NERSC</u>: (By mid-1995) The Center must conduct a study to determine which of the most heavily used engineering and chemistry applications codes will be ported to MPPs and which will be ported to SMPs by the owning institution or by the appropriate software vendor.

5. Porting applications supported by NERSC; (By late-1995) To port those applications which are both heavily used and maintained by NERSC, the Center must first determine which capability platform(s) to target. Those applications selected for parallelization (e.g. EFFI) should be parallelized using SAS to develop and test MPP porting tools.

6. Graphical User Interface: (By late 1995) A common GUI interface for most applications codes should be developed to facilitate use of the codes.

7. Parallel Mathematical Libraries Project: (By late 1995) This project will augment, if necessary, those parallel mathematical libraries provided by vendors. The project leader will conduct a search to determine which parallel libraries already exist and determine if they are appropriate for installation at or porting to NERSC. Finally, this project will ensure that serial versions of existing math libraries are provided on both SAS and the MPP to help with the parallelization of user codes. It may be necessary to hire or matrix staff to NERSC for this effort.

8. SAS: The Center must determine which supercomputer tasks (if any) should be moved exclusively to SAS to free up supercomputer cycles.

# Assimilation Environment

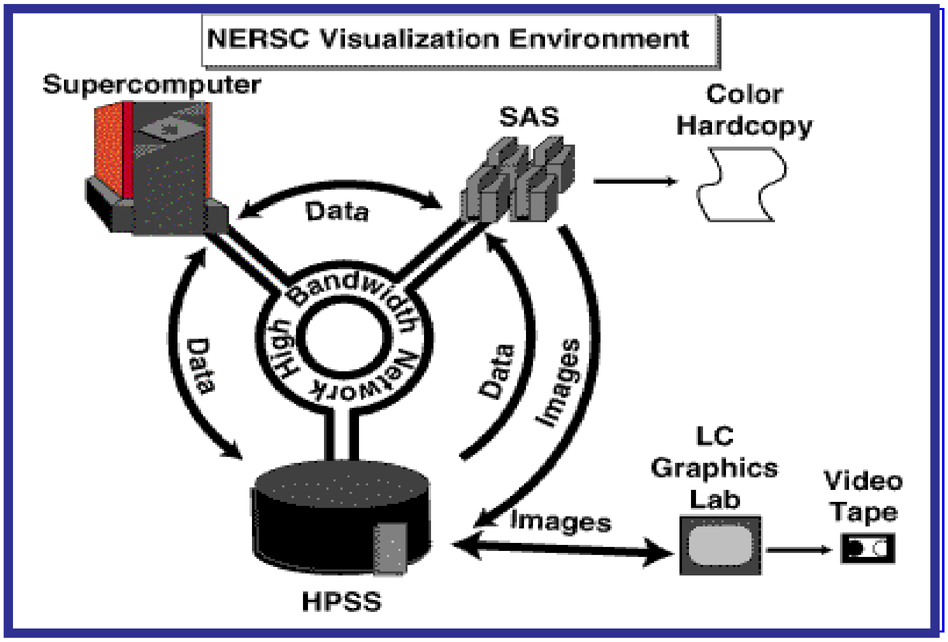
Every calculation generates data. How this data is organized so it can be assimilated by the human mind (which is currently not quadrupling in capability every three years) is the focus of the assimilation environment, the third element of the DCA triad. In the past, the primary focus of this effort has been to make it possible for the broad community to run any of a large number of packages, some home grown, others not, some involving the inclusion of arcane graphics calls within the simulation code, and others postprocessing data sets on supercomputing auxiliary servers.

All of this has worked passably well, until now. What has changed things is that the newer massively parallel computers can accommodate refined high dimensional physical systems and can integrate for long periods of time relative to the dynamics of the system modeled. In order to understand the simulated process, the physicist must be able to visualize a three-dimensional picture in time. How this rendering is effected cannot be achieved through a modest graphics package provided to the users with simple appended instructions.

Since the data generated will become increasingly voluminous, the researcher is forced to mine large stored data sets or to render on the fly, culling through data all the while. Either operating mode requires a far more unified production environment than exists at present at NERSC to work effectively. In other words, the rendering and graphics engines, the storage devices, the LAN, the WAN, and the high-performance computer must cooperate through a complex synchronization involving detailed protocols. To utilize this effectively requires human expertise.

The conclusion then is clear: *the assimilation environment can no longer be viewed as an independent appendage to the production environment. It must be an integral part of the UPE, or the data generated by the newer machines will become increasingly inaccessible. The following goals provide a feasible approach to this end.*

***Two basic modes of operation will be employed within the visualization environment: the first mode emphasizes post processing data placed in tertiary and secondary storage areas, and the second emphasizes real time coordination between the simulation as it is running and the rendering and visualization processes.***



Either of the two visualization operating environments that we propose can by represented by the graphic above.

- The first mode will be to store generated data within the High Performance Storage System (HPSS). The facilities of the Supercomputer Auxiliary Service (SAS) would then visualize that data. The resulting visualization images could then be sent to hardcopy devices for color output, to video equipment, or could be stored back on the HPSS. This mode will be necessary for users in the exploratory phases of their research, when visualization parameters are not well known.

- The second mode of operation in this environment will involve sending data directly from the supercomputer to SAS. In this mode, data will be directly visualized, and only the resulting images will be stored. In most cases, the image resulting from a visualization requires considerably less storage space. Direct visualization will enable users to visualize orders of magnitude more data than by post processing.

***There will be two levels of support and service provided by NERSC to the user community: baseline and advanced level support.***

- <u>Baseline Support Program</u>

  The emphasis in this mode will be to give the community a basic scientific visualization functionality that can be achieved without extraordinary effort on the part of the user. We list below the components of this service structure.

*Software:* It is essential that the number of software packages available and for which NERSC has consultative responsibility be kept to a small number and represent widely available, portable software. The graphics environment on the *vector supercomputers* will remain similar to its current state. The *newer computational environments* (the MPP or other multiprocessing systems) will not support locally developed and obsolete packages such as GRAFLIB or TV80. A strategy will be developed for moving codes from obsolete to supported baseline software package.

*Desktop functionality:* In order to provide continued support for graphics hardware, we must standardize on a certain set. We emphasize here two emerging standards to adopt at NERSC to effectively support our users: (1) *X windows* is recognized as the de-facto windowing system for UNIX platforms. It is essential that the Center make a concerted effort towards ensuring that 100% of our user base has X windows support on desktop machines. *(2) Postscript* may be easily classified as the de-facto standard for graphics output. NERSC should encourage remote users to obtain Postscript-compatible printers.

- Advanced Support Program

This approach will emphasize intense collaborations between NERSC staff and researchers intent on utilizing the full potential of the NERSC graphics laboratory and its resident experts. These collaborations will result in informative and striking scientific visualization results that will be used in publications and in other media to advertise the advances that can be made by computing at the highest end. Since, in all probability, the techniques developed will be generalizable, the fruits of these collaborations must be fed back into the general user base. In this way the entire community will benefit. We follow with some of the components of this program:

*Hardware:* As the volume of generated data to be visualized approaches terabytes, the Center must obtain a visualization system commensurate with those requirements. This should include a powerful graphics engine such as a multiprocessor ONYX with HIPPI attached to a high performance storage system (such as the mini-NSL). The complexity of the work done at NERSC, especially the requirement for 3D visualization, requires that we begin to examine the use of immersive technology (virtual reality), and this suggests the addition of a virtual reality board to the workstation. At a minimum, a high quality 2-D visualization medium (such as the “Wall”) must be made available.

High quality color hardcopy, as well as video animations, will be a critical component of any advanced visualization environment at NERSC. One option is to establish closer links to the LC graphics lab. This could be achieved by providing a HIPPI connection between the LC and NERSC graphics labs. This would benefit NERSC by providing access to advanced hardcopy and video tools without incurring the cost of obtaining or operating them.

*Software:* We plan to use off-the-shelf technology whenever possible. By placing an emphasis on using commercial or public domain software, we minimize development time and maximize the flexibility of the systems we develop. Commercial visualization systems such as AVS provide the power and flexibility needed to produce advanced visualizations. Packages such as AVS also provide the means to explore methods for post processing and direct visualization of data. Since AVS is available for lower-end systems as well as high speed systems, it provides a natural means for feeding the results of advanced work back into normal production, and also provides a pathway from standard visualization to move toward advanced techniques.

*Staffing:* It will be necessary for the graphics group members to combine efforts with the research and distributed computing groups. This will expand the range of expertise at NERSC and permit more collaborations with user clients without the necessity of adding graphics staff.

*Both efforts, Baseline and Advanced, must be links in a "closed loop", where advanced work is fed back into the baseline level, and the baseline level provides a direct path towards advanced visualization.*

### Milestones

1. Increase staff: (By mid- 19995) The current graphics effort at NERSC is composed of two FTEs, a staffing level sufficient for providing the Baseline Support Program only. NERSC should first hire (or assign) a programming technician for software installation. This person will be responsible for installing software on the supercomputing and SAS environments. This person will also be responsible for maintaining a database of installed software and contact points.

   Second, a new graphics staffer should be added to the effort. This person will support the baseline environment and encourage users to utilize increasingly sophisticated techniques as required by their evolving applications freeing existing staff for advanced visualization collaborations.

2. Broaden the graphics knowledge base: (By mid-1995 As visualization is truly a multidisciplinary field, it requires many areas of expertise. In the first phase, we will find members from most NERSC groups that will act as liaison for the visualization effort. Each member will focus on the aspects of visualization that pertain to their group, and will ensure that other members of their group are aware of these aspects.

3. Develop an expert system (by mid-1996) for answering questions regarding visualization. This system will be in the form of a searchable database for answering questions on visualization and the environment at NERSC. This system will provide the means for answering common questions. It will also be able to lead the user from starting points down pathways toward their visualization goals. When developed, this system will help not only in educating our staff, but will also be an invaluable consulting tool.

4. Acquire the hardware necessary for the Advanced Visualization Support Program. (complete by 1996, begins 1995) The Center should develop an acquisition plan which is cost effective and which will be in place before the upgrade to the FCM, and preferably in place at the time the PEP system is delivered.

5. Develop early collaborations: (Begins early 1995) This will involve finding users willing to try new techniques, and entering into pilot projects with them. Emphasis will be placed on publishing and disseminating results to NERSC users, so that the experiences gained may be fed back into our user base. Detailed case studies will be made available in the visualization expert system database, and as projects are completed, new users with more challenging problems will be sought.

6. X window capability: (Begun in 1994, complete by mid-1995) In order for this effort to be successful, the Center must ensure that NERSC staff and user clients have a minimal level of expertise in use of the X windows system.
   - The first step has already been taken through a series of *Buffer* articles.
   - NERSC should have a representative sample of desktop machines.
   - The graphics and distributed computing groups must take the lead.

# STORAGE

## Scope of Responsibility

NERSC must provide a broad range of storage services and facilities to DOE energy researchers throughout the United States. Centralized mass storage has been an integral part of NERSC since the Center began in 1974. The current archival storage system, the Los Alamos-developed Common File System or CFS, has provided centralized storage since 1987. CFS replaced an even earlier file storage system (FILEM) that evolved and served the Center over its first decade of operation.

The primary goal of centralized storage at NERSC has historically been to provide an archival service for users of NERSC supercomputers. However, due to the distributed nature of NERSC's computing community and the continued growth of ESnet, the strategic importance of non-NERSC machine access to NERSC storage has steadily increased.

## Areas of Emphasis

Computing is in transition. There are new games and new rules. New requirements and challenges now motivate the need for a new generation of storage systems with new capabilities. These requirements and challenges are driven by both the pull of evolving scientific research applications and by the push of advancing computer technologies.

To become broadly usable in wide-area computing, communication, and information infrastructures, storage systems must become invisible components supporting textual and graphic information, interdisciplinary scientific data, and multimedia applications. The interfaces to storage services need to provide uniformity across different computing platforms and computing environments. The services themselves must be intelligent, flexible, and enterprise-wide.

Scalability is critical. Future data set sizes, data transfer rates, and storage system name spaces are all projected to be much larger than current systems can support. Storage system topologies will become much more complex, with multiple storage systems located in different geographical areas. These systems will need to support logical single-system integration, and easy access by widely varying clients.

Storage systems will be required to live and evolve over decades. Such systems must be portable, adaptable to new applications, and accommodating to new storage hardware and software technologies. Use of standards for architecture, function, communications, security, and management will be a necessity.

NERSC has identified several strategic goals for storage. These goals fill holes in the Center's existing storage capabilities and improve our ability to offer state-of-the-art computational, networking, and storage resources to our users.

***Develop new storage management policies for NERSC that balance realistic and efficient utilization of available storage resources with voracious user demand. Consistent and well-understood methods of charging, quota enforcement, purging, and local disk management through these new policies is a major goal.***

***Increase the level of reliability, recoverability, and robustness of the storage environment to improve data management, data storage, and data access.***

***Provide scalable storage services and facilities that have potential to reduce the ever-widening gap between storage device mechanical latencies, and processor, memory, and network electronic-and-optical technologies.***

***Provide new interfaces or access to storage services for users that want or need "local file system" views. Remove requirements that force explicit knowledge about remote or archival facilities for those users that don't want to know about anything but their local environment.***

***Provide adequate facilities for user-directed archiving or user-directed export (physical removal or mailing) of user-owned data. Similarly, provide adequate facilities for system-directed archiving or export of system-owned data. Provide adequate automated backup capabilities for home directories and work spaces.***

***Provide support for heterogeneous client environments, particularly in the form of accessibility from non-NERSC hosts, workstations, and personal computers in conformance with NERSC storage management and access policies.***

***Together with the above goals, provide an ability to support and manage the following dichotomous views of storage, as appropriate for particular users:***

1. ***transparent vs. explicit knowledge of storage resources***
2. ***file-oriented vs. data-oriented user interfaces***
3. ***user-owned vs. system owned storage volumes***
4. ***locally-managed vs. remotely-managed administrative domains***
5. ***high-performance vs. low-performance classes of service***
6. ***uniform charges vs. variable charges***
7. ***uniform quotas vs. variable quotas***

There are several focus areas or milestones for storage that need to be pursued. These focus areas address one or more of the previous strategic goals for storage. By quantifying issues, analyzing current and future technological possibilities, and suggesting tactical approaches, work in these focus areas will help realize NERSC's storage goals.

### Milestones

1. Formulate NERSC hardware and software acquisition plans. (Begun in 1994) This area is already underway. See Appendix 1 for a summary of planned acquisitions

over the next three years for centralized storage.

2. Investigate the possibilities of integrating AFS disks and servers with HPSS. (Begun in 1994) This area is already underway. A possible collaboration with the University of Michigan to integrate DFS facilities with HPSS is being explored.

3. Study other available or widely-used file system technologies as appropriate storage interfaces for NERSC (e.g., FTP, NFS, VFS, AFS, DFS). Examine the implications of new Internet navigational tools such as Mosaic.

4. Study available or widely-used backup and restore (BAR) applications and automated data migration facilities such as CRAY DMF and products based on the Data Migration Interface Group (DMIG) interface specification.

5. Identify other necessary and/or desirable storage services that will be built on top of an HPSS/file system infrastructure (e.g., multimedia document archives, X-Window directory structure browsing tools, POSIX file system support etc.)

6. Identify necessary and/or desirable storage management policies for migration, purging, backup, charging, quotas, etc. and the policy management applications that will need to exist in order to make effective use of such policies.

7. Develop appropriate decision models for storage by investigating past storage usage statistics, anticipated usage behaviors, current and future network bandwidth characteristics, and gigabytes generated per teraflop estimates.

8. Investigate the probable implications of increased use of storage archives for large scientific and technical data management applications, as well as possible collaboration opportunities (e.g., LC Intelligent Archive, Sequoia 2000, ARPA).

9. Investigate the current and future use of standardized media, data structures, data formats, metadata, and transaction management facilities and their implications on storage systems and storage management at NERSC.

10. Investigate NERSC-specific class-of-service issues, including parallel I/O and potential concurrent access problems arising from the responsibility to support thousands of highly distributed users.

11. Identify possible new markets for NERSC storage service beyond the traditional supercomputer user base. Should we attract new customers for NERSC storage? Should storage use be limited to support of NERSC computing infrastructure?

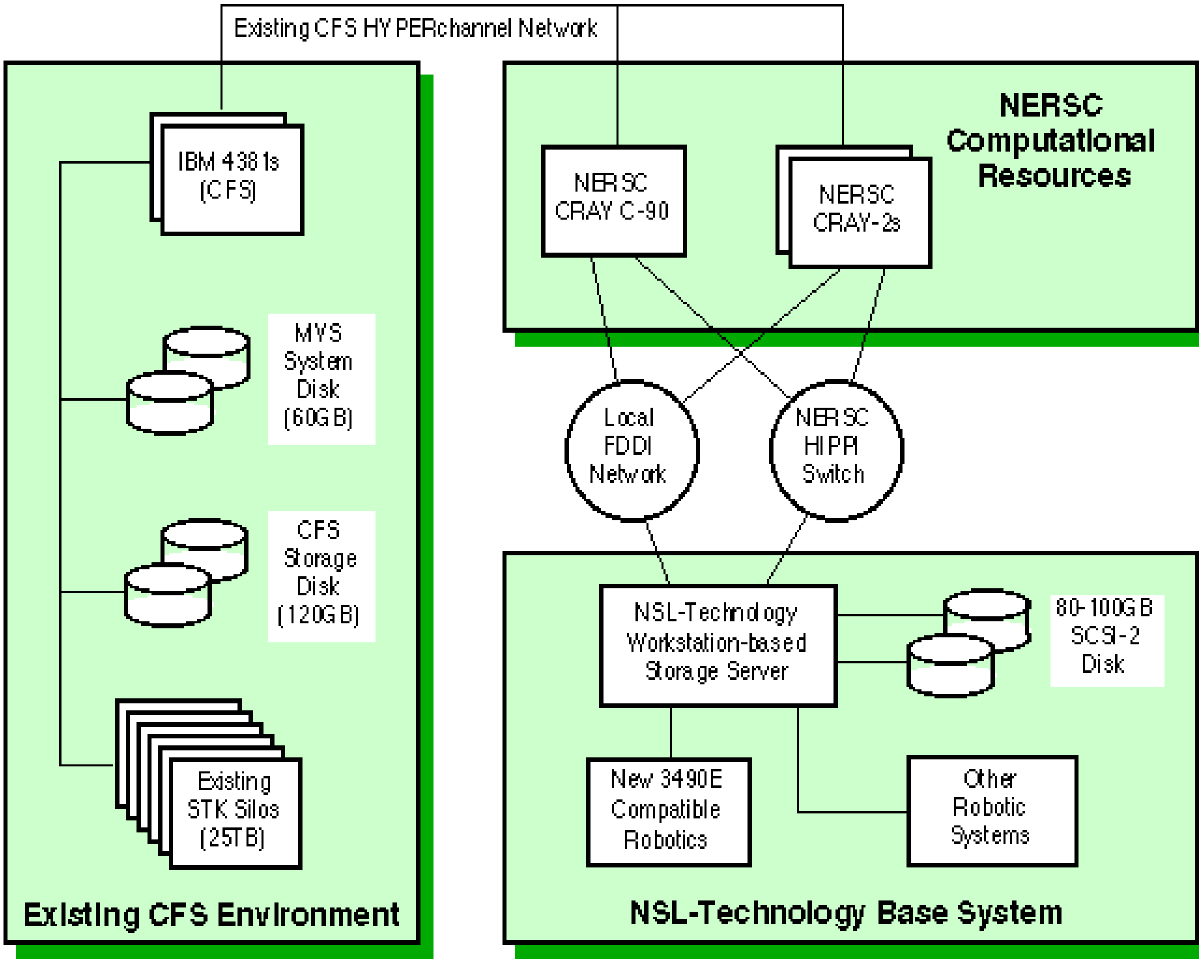
# Appendix 1: NERSC Storage Acquisition Plans

NERSC has studied several approaches  for upgrading, supplementing, and replacing the Center's existing centralized storage system environment to meet our strategic goals for storage. The implementation plan for upgrading NERSC's centralized storage facilities addresses three distinct phases of storage system technology acquisitions:

**Phase I** establishes a small base-level storage system based on proven National Storage Laboratory (NSL) hardware and software technologies that will run in parallel with the existing CFS system

.

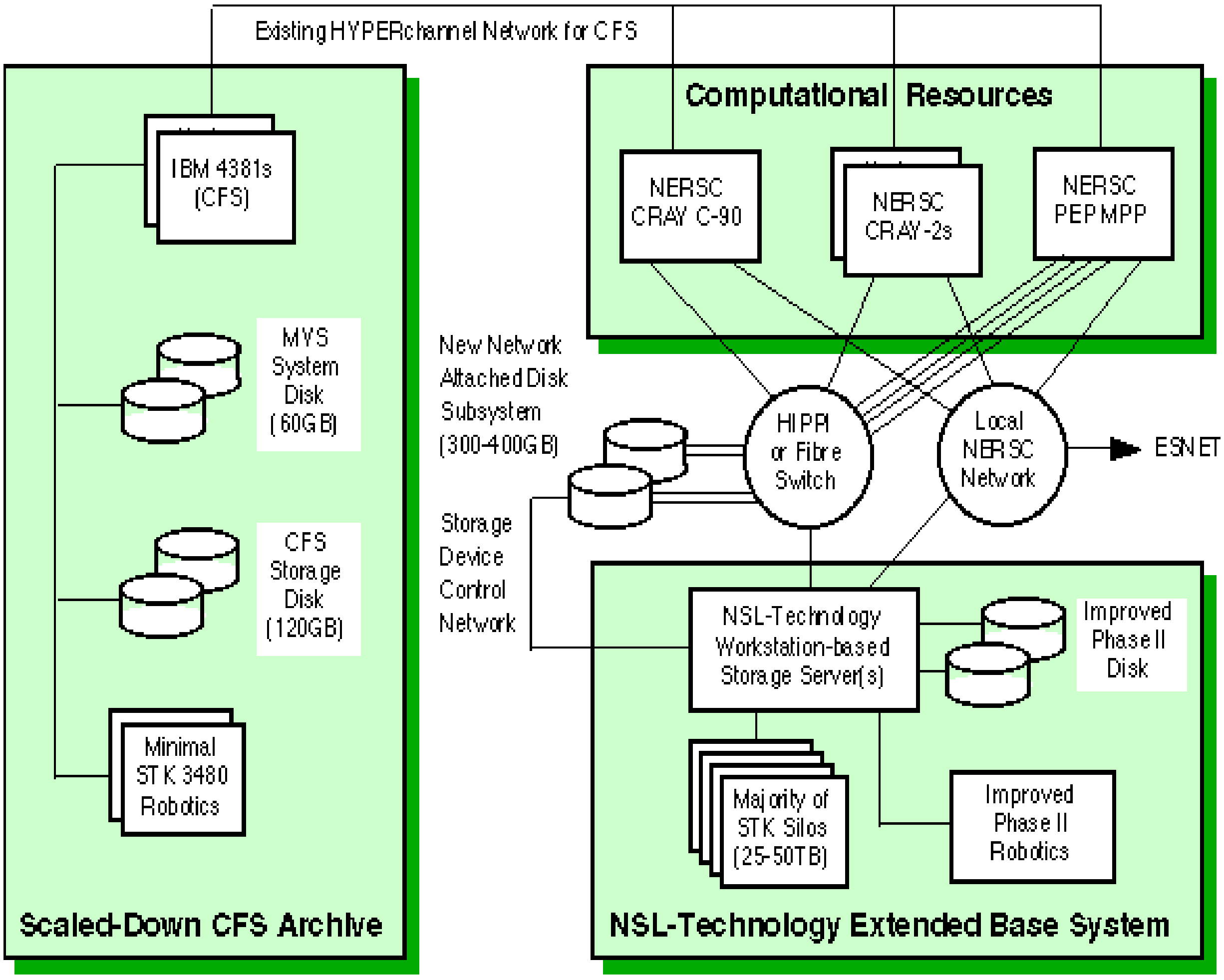
## NERSC Phase I Storage Environment



**Phase II** establishes a large, interim production storage system built over the Phase I base-level infrastructure through additional high-speed network interface connections and network-attached array disk subsystems,  and coupled with a controlled scale down of the CFS environment

.

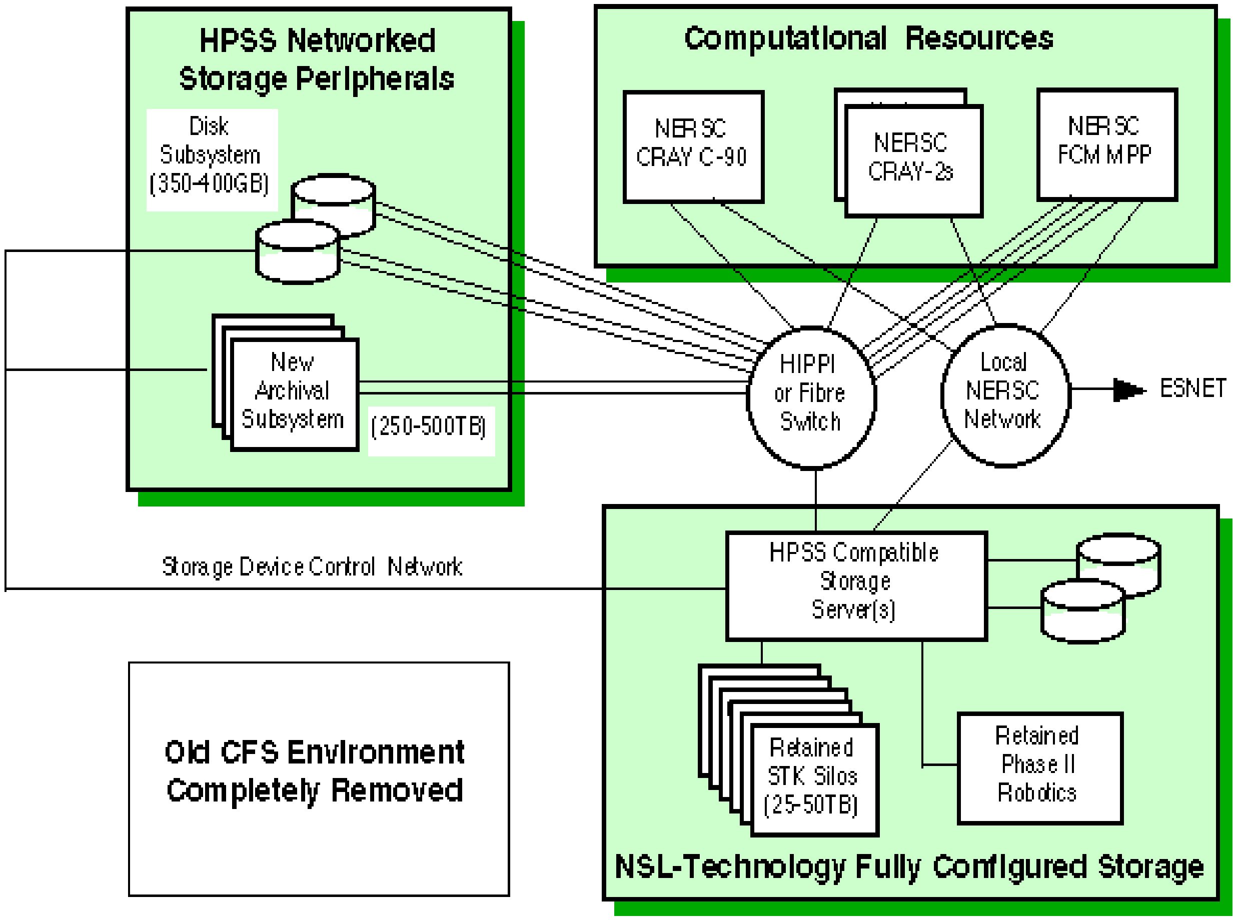
## NERSC Phase II Storage Environment


Existing HYPERchannel Network for CFS
Computational Resources
NERSC CRAY C-90
NERSC CRAY-2s
NERSC PEP MPP
IBM 4381s (CFS)
MVS System Disk (60GB)
CFS Storage Disk (120GB)
Minimal STK 3480 Robotics
Scaled-Down CFS Archive
New Network Attached Disk Subsystem (300-400GB)
Storage Device Control Network
HIPPI or Fibre Switch
Local NERSC Network
ESNET
NSL-Technology Workstation-based Storage Server(s)
Improved Phase II Disk
Majority of STK Silos (25-50TB)
Improved Phase II Robotics
NSL-Technology Extended Base System

**Phase III**, the final phase, addresses transfer or removal of all remaining CFS storage devices, and addition of a new high-speed, high-capacity archival subsystem. This phase also includes the likely replacement of initial NSL software systems with High Performance Storage System (HPSS) software. It may be possible to deploy HPSS in a production environment prior to Phase III. NERSC feels that an HPSS-based environment is much preferred over continuation of Phase I base-level software. NERSC's belief is that HPSS, with its scalable, parallel I/O architecture, will be much better equipped to handle the large data transfer and storage requirements placed on it by the introduction of new massively parallel machines

.

## NERSC Phase III Storage Environment


HPSS Networked Storage Peripherals
Disk Subsystem (350-400GB)
New Archival Subsystem
(250-500TB)
Computational Resources
NERSC CRAY C-90
NERSC CRAY-2s
NERSC FCM MPP
HIPPI or Fibre Switch
Local NERSC Network
ESNET
Storage Device Control Network
HPSS Compatible Storage Server(s)
Retained STK Silos (25-50TB)
Retained Phase II Robotics
NSL-Technology Fully Configured Storage
Old CFS Environment Completely Removed

# ADMINISTRATION

## Scope of Responsibility

To create and maintain a consistent and well integrated management environment across major NERSC services.

## Area of Emphasis

Traditionally, the Center has concentrated on providing a number of discrete services which were complementary but fundamentally unintegrated. This was so because network technology (LAN and WAN) was limited, and because it made some sense to treat computational platforms as self-contained, autonomous systems. That period is over.

Soon NERSC will have a number of diverse capability platforms on the floor. There will also be, in addition to a heterogeneous SAS environment, a variety of secondary and tertiary storage devices, complex visualization hardware, extremely fast and high bandwidth interconnects, and user clients with significant local capabilities who want to feel at home no matter which system they are currently using, local or remote. What has happened is that the number of types of services has increased, as has the degree of specialization of the components out of which these services are constructed. Unless these services can be glued together transparently, the scientist will be unable to utilize the potential of the hardware that has been made available. This “glue” is what we have called the Unified Production Environment (UPE).

As the researcher employs various NERSC services to facilitate progress through all phases of a computational research project, the sense should be that only one environment exists. After logging in, the researcher will see a shared file system regardless of which service is used. At a later stage of the development of the unified environment, the user will have the tools required to effectively utilize the resources through distributed batch and interactive computing. We can view this as a *unification of centralized services.* Ultimately, the Center will seek to integrate the remote user's local environment with the centralized NERSC environment. This could be viewed as a *unification of distributed services.*

The emphasis of this chapter is that the UPE must include an integrated management environment with a consistent management interface for each service provided in the UPE.

## Background

The computing world has been evolving toward what is commonly viewed as a consistent and well integrated environment based upon UNIX. One discovers upon closer examination that very significant incompatibilities exist within what is viewed as UNIX and that providing a well integrated environment comes only with a great deal of effort. Even systems developed for distributed computing, such as the Kerberos authentication server, exist in several incompatible versions. Operating systems developers choose one version of the Kerberos authentication server and, by virtue of that choice, can only form a well integrated environment with computer vendors having like-minded developers. Fundamental shortcomings of UNIX, such as system security, have led to the development of incompatible and inconsistent system enhancements by each computer vendor.

Until such time as UNIX evolves into a consistent and well integrated environment, the administration of these systems will be a challenging vocation. Systems administration is different on each computer system and NERSC has several groups performing systems administration on the various systems. Some flavor of UNIX is the operating system on most of these computers, although exceptions persist. Given that there is some degree of commonality in the operating systems, it would benefit NERSC to coordinate the administration of the computers. In fact, a high level of coordination is essential to the mission of a consistent and well integrated management environment. Secondarily, a high level of coordination could increase the productivity of existing staff and increase their breadth of understanding through improved communications and exposure to additional systems.

To achieve this vision, we will need to establish goals and the milestones required to achieve those goals. In the following pages we do this for the major administrative areas of activity at NERSC.

## Integrated Systems Management

***In order to provide an integrated management environment, NERSC must strengthen the team concept among the systems administrators and make consistent the tools and policies they utilize.***

NERSC needs mechanisms to ensure timely and consistent management of its systems and services. Such mechanisms and/or tools should enhance the productivity of NERSC administrative staff. System management is performed relatively independently in several of NERSC’s groups. Integrated management will require a higher level of cooperation between those performing system management.

**Milestones**

12. NERSC systems administration team: (By mid-1995) Establish a project or team responsible for systems administration throughout NERSC. The team leader will work from a clear charter and use strong leadership to effect the communication and coordination required. The team members will come from various groups at NERSC.

13. Integrated trouble ticket systems: (by mid-1996) Problem reports for both ESnet and NERSC are tracked centrally in order to insure a prompt response and eventual problem resolution.

14. Uniformity: Develop a mechanism to enforce consistent policies, uniform security, and uniform administration of NERSC systems and services including licensed application software.

## Security

***Besides ensuring that our customers are properly authenticated (covered in a separate item), we need to provide sufficient protection of their data and ensure that NERSC systems are sufficiently protected against misuse.***

**Milestones**

1. Foreign nationals: (By mid-1995) Identify NERSC customers who are foreign nationals and restrict their access to sensitive data. This will involve CUB, archival storage and other NERSC services.

2. Collaborations: Identify possible collaborators for computer security issues

3. Vendor supported machines: Ensure sufficient system security on computers located at NERSC, but not directly administered by NERSC (i.e. vendor supported machines).

4. Identify top security threats. (By mid-1995)
   A. Establish a team to identify major threats. LLNL has such a team, which includes NERSC staff. They have already completed this milestone and have produced a collection of recommended security measures (see **Appendix 1**).
   B. Develop a plan to reduce or remove these threats.
   C. Periodically review security threats.

## Authentication

***NERSC needs to provide a high level of security for its customers without significantly sacrificing ease of system use***

The weakest link in the system security now is transmission of unencrypted multiple-use passwords across networks. Single-use passwords can substantially reduce the risk of unauthorized access and must be supported by NERSC. We can also provide for greater ease of system use with improved security through distributed authentication servers, such as Kerberos. This will provide our customers with the ability to perform work on multiple NERSC servers after a single NERSC authentication.

**Milestones**

1. Single-use passwords: (by late 1995) Provide support for single-use passwords based upon smart cards and/or s/key. We must convert to single-use passwords for NERSC systems and support personnel with special privileges. Our customers can purchase the required hardware or software, as they see fit, but NERSC must provide the authentication support.

2. Login aliases: (by late 1995) Enable users to select a “real name” alias as their NERSC login name.

3. Single NERSC login/authentication. (by late 1995)
   A. AFS could provide this capability for now. We will utilize DCE security service (Kerberos 5) when appropriate.
   B. Kerberize telnet, ftp, etc.
   C. Combine a one-time password system with KDC.
   D. Kerberize or remove remote commands.
   E. Determine how to support customers with “dumb” terminals.
   F. Provide seamless ticket forwarding.

4. Distributed authentication: (by early 1996) Provide cross-realm, Kerberos based authentication (i.e. authentication between computer centers).

## Banking

***NERSC needs to provide to its customers a highly reliable allocation and accounting system spanning all of its services.***

**Milestones**

1. T3D: (by early 1995) Port CUB to the Cray T3D.

2. SAS: (by late1995) Port CUB to SAS/HP/Solaris.

3. Multiple repositories: (by early 1996) Allow a single user to have access to multiple repositories.

4. Archive quotas: (by early 1996) Provide archive quota support for NSL/HPSS.

## Distributed Computing

***Expand the AFS service to all NERSC customers and transition to DFS.***

A distributed file system is the most important facet to distributed computing. This will entail support of distributed file systems at NERSC, both the clients and servers. NERSC must also encourage use of distributed file system by its customers to enhance their productivity, sharing of files, and reduce system overhead. A distributed file system supplies the mechanism for providing centralized authentication, shared home directories, and X security via the shared home directories. Shared home directories will eliminate duplicated data, user effort in moving files, inconsistent versions of files as well as facilitating the use of the optimal compute server for each task and sharing of data.

***NERSC must also provide the tools required to effectively utilize all resources through distributed batch and interactive computing.***

**Milestones**

1. AFS/DFS clients: Provide an AFS and/or DFS client on every NERSC compute server for which one is available.

2. Translators: For systems which do not support AFS or DFS, support an AFS/NFS or AFS/AFP (Macintosh) translator.

3. DFS servers: (by early 1996) Transition distributed file system support from AFS to DFS when appropriate.

4. ER-wide file system: (by mid-1996) NERSC will participate in ER committees to establish a common file system across participating sites.

5. Archival storage: (by late 1995) Integrate the distributed file system with NERSC archival storage.

6. Shared home directories: Support shared home directories on every NERSC compute server.

7. Distributed batch: (by mid-1996) Provide the ability to submit and monitor batch jobs from any NERSC compute server.

8. Distributed interactive computing: (by mid-1996) Explore potential load sharing of interactive tasks between systems.

## Licensed Software Administration

***Provide NERSC customers with convenient and cost effective access to licensed, third party software applications that satisfy the customers' programmatic requirements.***

We include here both applications that run on supercomputers and supporting applications that run on workstations. The goal is for NERSC to maximize the effectiveness and flexibility of providing licensed software to our customers while minimizing the administrative burden.

**Milestones**

1. Central repository: (by mid-1995) Tapes, disks, and instructions for installing software should be stored in one location.

2. Organization of binaries: (by mid-1995) Records shall be kept indicating where (e.g. what directories on which machines or AFS) the software is installed.

3. Documentation: (by late1995) All licensed software should have at least a man page or some other consistent starting point for finding further documentation. Not only should the documentation cover usage of the applications, it should describe licensing arrangements and the steps a NERSC customer must take to use an application under a floating license.

4. Usage statistics: Using any feasible mechanism, usage records should be kept of licensed software for the purpose of analyzing the adequacy of the number of licenses versus demand and the cost of licenses versus total usage.

5. License manager: (late 1995) Explore the use of a license manager program to check-in and check-out floating licenses to customers to run either at NERSC or at the customer's site.

6. Technician: (by late 1995) NERSC should hire a programming technician to perform the necessary work to achieve the aforementioned goals. The specific tasks in the job description are listed in Kirby Fong's paper, ”Mission in Licensed Software Administration” (see **Appendix 2**).

# Appendix 1: LLNL Guidelines on Levels of System Protection

Six levels of protection for machines connected to the Internet.

**Protection Level 1** - All systems need this level of protection:

Protection Goals

1. Plug the known holes in the operating system.
2. Configure and maintain the system to only give rights and privileges to those who should have them.
3. Be able to recover corrupted or destroyed data.
4. Do not allow hackers to grab the root password as it transports across the net.
5. Limit access to only those remote machines that should reasonably have access.
6. Have the computer under the guidance of someone knowledgeable in computer security.
7. Make sure the computer's users are aware of computer security issues.

Recommended actions

1. System is running an up-to-date OS with all security patches.
2. System under good/defensive system administration.
3. Data is backed up to an off-line media on a regular basis.
4. Use one time passwords (S/key or Enigma Logic) for root account.
5. Use TCP Wrappers.
6. System security is monitored and guided by a trained CSSO.
7. Users trained on computer security relevant to their machine(s) and the sensitivity level of the data on them.

**Protection Level 2**

Protection Goals

1. Prevent hackers from grabbing user passwords as they transport across the net.

Recommended actions

1. Take all Level 1 actions plus
2. Use one time passwords (S/Key or Enigma Logic) for all users.

**Protection Level 3**

Protection Goals

1. Stop hackers from looking at sensitive data files on disk.

Recommended actions

1. Take all Level 1 and 2 actions plus
2. Install suitable encryption tools.
3. Encrypt all sensitive data files on disk - this is the user's responsibility.

**Protection Level 4**

Protection Goals

1. Prevent hackers from snooping information as it transports across the net.

Recommended actions

1. Take all Level 1, 2, and 3 actions plus
2. Encrypt all data that goes out onto a network - this is the user's responsibility.

**Protection Level 5**

Protection Goals

1. Deny Internet access to machines with sensitive data.

Recommended actions

1. Take all Level 1, 2, 3, and 4 actions plus
2. Isolate machine from the Internet with a Firewall and monitor the firewall computer.

**Protection Level 6**

Protection Goals

1. Guarantee that outside attackers can not access sensitive data.

Recommended actions

1. Disconnect from the Internet

# Appendix 2: Licensed Software Administration

## MISSION

The mission of licensed software administration is to provide NERSC customers convenient and cost effective access to licensed, third party software applications that satisfy customers' programmatic requirements. This includes both applications that run on supercomputers and supporting applications that run on workstations.

## VISION

This section concentrates on supporting third party software applications such as graphical pre- and post-processors, mathematical and data analysis tools, code development and maintenance tools, and documentation tools. These are highly interactive tools best run on a workstation rather than on a supercomputer. While the Supercomputing Auxiliary Service provides a place for these tools to run, it would be better if customers could run them on their own workstations for better interactive response and less traffic loading of the network. Heretofore this meant customers would have to buy and administer their own software licenses for their own computers; however, recent advances in license management technology now permit the following:

- NERSC can purchase and administer floating or network licenses for the necessary applications.
- NERSC customers do not need to sign license agreements or perform any purchasing actions at their own sites.
- NERSC customers can download and keep executable versions of the applications on their own computers.
- The application runs only if it can check out a license from a license manager program running at NERSC. The license is checked in when the application terminates.
- Through the administration of a license options file, NERSC can control which users or hosts are allowed to check out licenses. Permission to check out licenses would be granted only by request of the Principal Investigator and can easily be revoked in cases of abuse. (Since all users of SAS are authorized NERSC customers, all users logged into SAS will be allowed to use all software there without any special request by the Principal Investigator.)
- Since the license manager keeps a log of who checked out a license and when they checked it in, it is possible for NERSC to levy a CRU charge based on the amount of wall clock time a license was used. This discourages wasteful use of licenses.
- NERSC would acquire not only the versions of applications for the SAS platforms (HP and Sun Solaris) but would also acquire versions for other platforms commonly used by its customers (e.g. SGI, IBM RS/6000, DEC Alpha OSF/1, etc.)

Not all licensed products contain the floating license mechanism just described, but about half the tools already on SAS do. We believe this vision for administering licensed software maximizes the benefit-to-cost ratio to OER while providing researchers the tools they need on the most appropriate platforms.

## GOALS

The vision described in the previous section addresses specifically the SAS oriented licensed software. Our goals for administering licensed software, however, apply in many instances to the licensed software on the supercomputers.

- Tapes, disks, and instructions for installing software should be stored in one location.
- Records shall be kept indicating where (e.g. what directories on which machines or AFS) the software is installed.
- All licensed software should have at least a *man* page or some other consistent starting point for finding further documentation. Not only should the documentation cover usage of the application, it should describe licensing arrangements and the steps a NERSC customer must take to use an application under a floating license. See the document on Production Development Services for plans on NERSC documentation.
- Using any feasible mechanism, usage records should be kept of licensed software for the purpose of analyzing the adequacy of the number of licenses versus demand and the cost of licenses versus total usage.
- NERSC should hire a programming technician to perform the necessary work to achieve the aforementioned goals.

## SPECIFIC TASKS

This section lists various tasks and responsibilities for the programming technician in addition to achieving the goals described in the **GOALS** section.

- The technician will need to work primarily under the direction of the SAS system administrator but will also have to coordinate his or her work with the AFS administrator, operations staff, Applications Software Group members, Manager of Contract Software, and others. The work includes installing software applications, testing software installations, installing documentation, verifying accessibility of software and documentation, and conforming to NERSC policy on the announcement and release of new software. Such software may be on the supercomputers as well as SAS. This task includes staging in files from distribution tapes and disks.
- The technician must be responsible and careful enough to be entrusted with root access since the installation of software, license files, license managers, and options files may require privileged access.
- The technician must become knowledgeable in the interactions of various versions of license managers and their license files and will install multiple products in a manner that avoids mutual conflicts. This means the technician works with the SAS system administrator to assure the default login environment includes access to all the licensed products.
- The technician is the principal point of contact for NERSC customers who need assistance in setting up a licensed application to run on a remote workstation. The technician advises customers on what files are needed, where to get them, where to put them, and how to set up environment variables correctly. It may mean advising customers whose AFS cells are authenticated to the NERSC AFS cell how to use the application without explicit downloading.

- The technician administers the license and options files to add and delete remote computers on which floating licenses may be used.  The technician periodically saves the license log files and runs license management report generators.
- A one time assignment is to investigate FLEXadmin and FLEXwrap as potential systems for facilitating the administration of licenses and the managing of software applications in accordance with previously mentioned goals and tasks. (The FLEXlm License Manager is embedded in many of our current applications.) A continuing assignment for the technician is to seek and evaluate new tools that could increase his or her productivity in administering licensed software.

# LOCAL AREA NETWORK

## Scope of Responsibility

The mission of the local area network at NERSC is to deliver data as required by the various systems attached. These systems are the base platforms on which the user services are built. The requirements for data movement are initiated by the 4500 users of the NERSC computing and storage resources. The characteristics of these requirements range from very large, long duration file movement, to bursty, highly responsive interactive access. The NERSC LAN must address each of these requirements appropriately.

## Area of Emphasis

The NERSC local area network provides the communications links between the various systems on which the user services are built. In this critical role capacity is a key. A bottleneck in any of the segments will have far reaching impact on the overall services to the 4500 users. The entire local area network at NERSC needs to provide support for an array of platforms including: High Performance Vector, Scalar, and Massively parallel Computing; Large file storage systems; Auxiliary Data Servers; and Desktop PC, MAC, or Unix Workstations. Economically connecting this array in turn requires a variety of media types.

The High Performance Computing and Communications Initiative identified a National Research and Education Network (NREN) program which has focused attention on advancing the state of the communications infrastructure. Its goal is a nationwide 1G bps capability at the end of five years. In line with this program the Department of Energy and NERSC through the Energy Sciences network (ESnet) have embarked on a five year plan to upgrade from 1.5M bps through 622M bps to be prepared for the 1G bps NREN step to follow. As this takes place over a fairly aggressive time scale (without a matching local area thrust) it is easy to foresee the wide area bandwidth continuing to exceed the local area capacity at NERSC.  In addition, the next generation of scientific applications is expected to take advantage of the emerging increase in bandwidth with low predictable latency being deployed as part of the HPCC's NREN program.  A missing link in the ability to effectively use these applications is a local distribution network with similar or better characteristics to the wide area.

In recent months the wide area bandwidth connected to NERSC has exceeded 140 Mbps while the fastest local segment is only 100 Mbps. When the external network capacity meets or exceeds the local segment capacity, a single external system will be capable of serious degradation of local communications, and thereby NERSC services as a whole. This situation will continue unless serious emphasis is directed at the problem.

Historically the local networking at NERSC has evolved out of component availability rather  than been directed by an overall plan. This document will define a plan for the near term to bring the environment up to speed with its current wide area connections and into line with the state-of-the-art.  It also lays out a strategy, with the goal of maintaining the evolution beyond these first steps, to increase the capacity by an order of magnitude every 2 years.  Technology will march on, and to provide a world class computing environment NERSC will need to keep the local network up to the task.

## Background

The Local Area Networks continue to play a vital role in NERSC's ability to provide the services customers expect. The last installment [1] of this saga, addressed the need to move from the technology of the early 1980's to something more current. Looking back, the accomplishments of the last three years were, completion of outlined phase 1, and initial steps on phase 2.  The Ethernet concentrators are in place as well as an FDDI ring connecting the high end systems in the machine room of 451. While these are modest improvements, they are significant steps forward.  This document will update and outline a strategic direction in addition to a short term implementation plan.

The supercomputer and storage environments at NERSC continue to grow in capacity and capability. Added to this load are the emergence of baseline distributed computing services and low cost, powerful desktop equipment. In this environment managing growth becomes a significant issue. Recent traffic statistics in the wide area show a doubling of volume on the order of every 6 months.  This translates into more than an order of magnitude increase every 2 years. In addition, the pace of global technology improvements continues to increase.  Making use of these improvements, the current end systems will be starved for data waiting for access to the shared media Ethernet and FDDI.

Beyond managing the growth, just keeping tabs on the diverse technologies used presents a challenge. The traditional mode of management through portable diagnostic tools breaks down when problems span differing technology segments. It also ignores routine data collection to detect patterns or historical growth. A robust management system is required.

A new area of growth is travel and home computing. Creating a useful subset of an office environment at home or on the road is a large task.  The graphical user interfaces obtaining heavy use today are bandwidth consumers to a point that access speeds considered appropriate for large groups of users just 10 years ago, are now becoming bottlenecks to a single user. Initial steps to address this need were installation of a pool of analog dial modems at 14.4 kbps, and a limited quantity of digital services at switched 56 kbps or 128 kbps ISDN. More effort will be required.

[1] NERSC  LAN  EVOLUTION  Tony Hain  5/12/91

## Architecture

The collection of resources NERSC manages are shared among 4500 researchers pushing the state of the art in computing and scientific collaboration. Therefore NERSC's local network needs to take advantage of high rates to move large sets of data while not locking down on a single user.  The basic architecture is a multi-level hierarchy with performance appropriate to the systems attached. This architecture is relatively independent of actual interface speed, with the exception that each level provides approximately an order of magnitude more capability than the previous. New technologies are integrated at the top in a push down approach extending the useful lifetime of each generation.

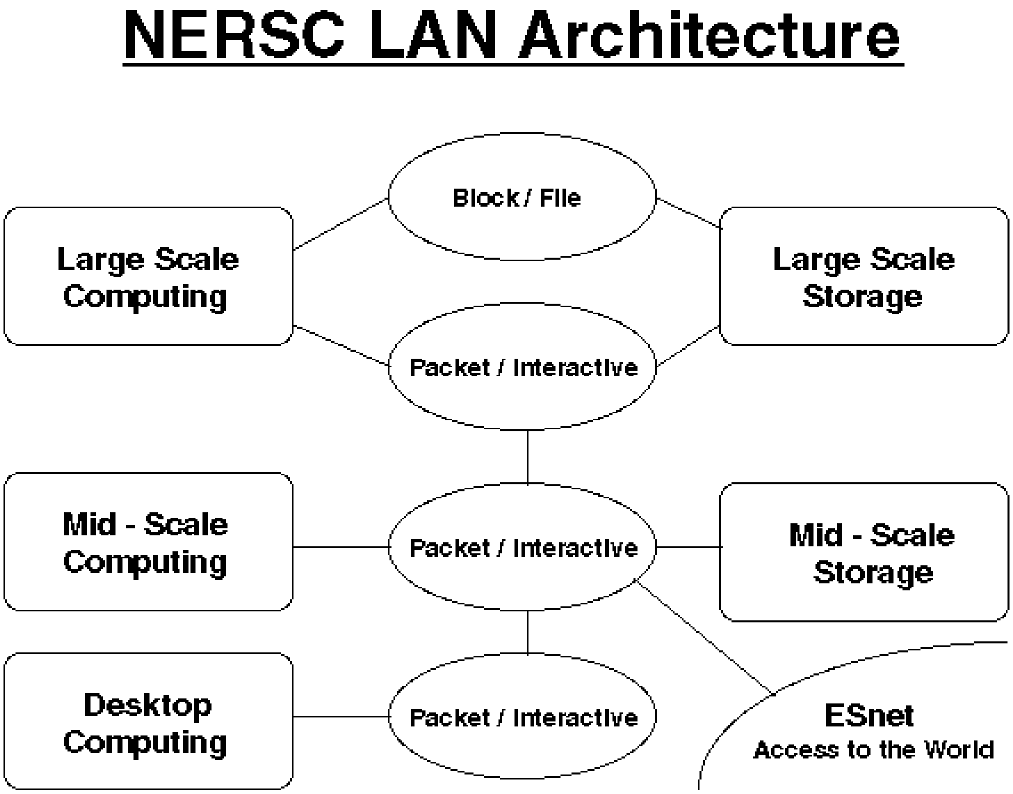


Fig. 1 -  NERSC LAN Generic Architecture

To address both multiple system types and high availability, the overall model is a multi-segment structure where the performance of each segment is matched to the performance needs of the attached systems. The multiple segments limit single points of failure as well as addressing specific needs. For example, an implementation that works well for file storage will present unacceptable delays to general purpose interactive access, while an implementation optimized for general purpose access will lower the effective throughput for file movements. Moreover, while there is a need for high performance on one hand, there is a need for cost effectiveness on the other. As these are generally conflicting requirements, several media types are required to support the mix.

The highest performance tier is that attaching the large file storage and high performance computing systems. The separate segments shown here are optimized (including the operating systems and device drivers) for the tasks of large file movement and general purpose access. At any given time this separation may require different technologies to meet the characteristics of the differing demands. Both segments will need comparable capacity as contemporary applications make use of models of ever larger complexity, through animated graphical displays requiring correspondingly large datasets from storage.

At the second tier the auxiliary computing and storage systems attach to a segment which is oriented to support interactive access, and a bursty file system. The wide-area interface connects here due to its comparable performance level, and to isolate the top tier from any unnecessary external load. Continuous availability is crucial for this tier in its pivotal role.

The lower tier provides the distribution to the desktop systems, including those traveling and at homes. Its primary characteristics are low interface cost and radial distribution wiring. Being at the end of an evolutionary cycle, this tier will see limited additional investment beyond what is required to expand the port count to cover the greater number of systems attached.

Management of the network complex will use SNMP tools of the same type used to manage the Energy Sciences Network. This will allow a high level of support without significant additional expenditure or effort in training. The management system will provide a graphic display of the current state of the attached systems, as well as data collection from the network nodes measuring traffic characteristics. Additional diagnostic tools are required to deal with each of the segment technologies implemented. These tools should be integrated with the management station, or available to multiple users via an X-window interface when integration is not possible.

## Implementation

Currently interconnection of the large scale computing and file storage systems is limited to 50 Mbps aggregate by the Hyperchannel technology. Individual transfers are limited to the 24 Mbps channel rate of the CFS servers.

The server class systems are connected with each other, the large scale computing systems, and the outside world with an FDDI ring network. The ring provides both aggregate and individual transfer capacities of 100 Mbps.

Desktop access is supported via Ethernet concentrators which provide thinwire connections to the office. The aggregate to each wing (avg. 20 offices) is 10 Mbps. These wing Ethernets are connected to the FDDI ring by a pair of routers.

Home and travel access is provided via 14.4 kbps analog dial modems connected to the Desktop tier. There are currently 40 modems providing a range of access services including, generic terminal, Appletalk, SLIP, PPP, and X-remote.

Current network management is a combination of monitoring from the ESnet management stations and ad-hoc portable tools.

The evolutionary development of the NERSC LAN has been in the general direction of the architecture goal. Its current implementation is shown in Fig. 2.

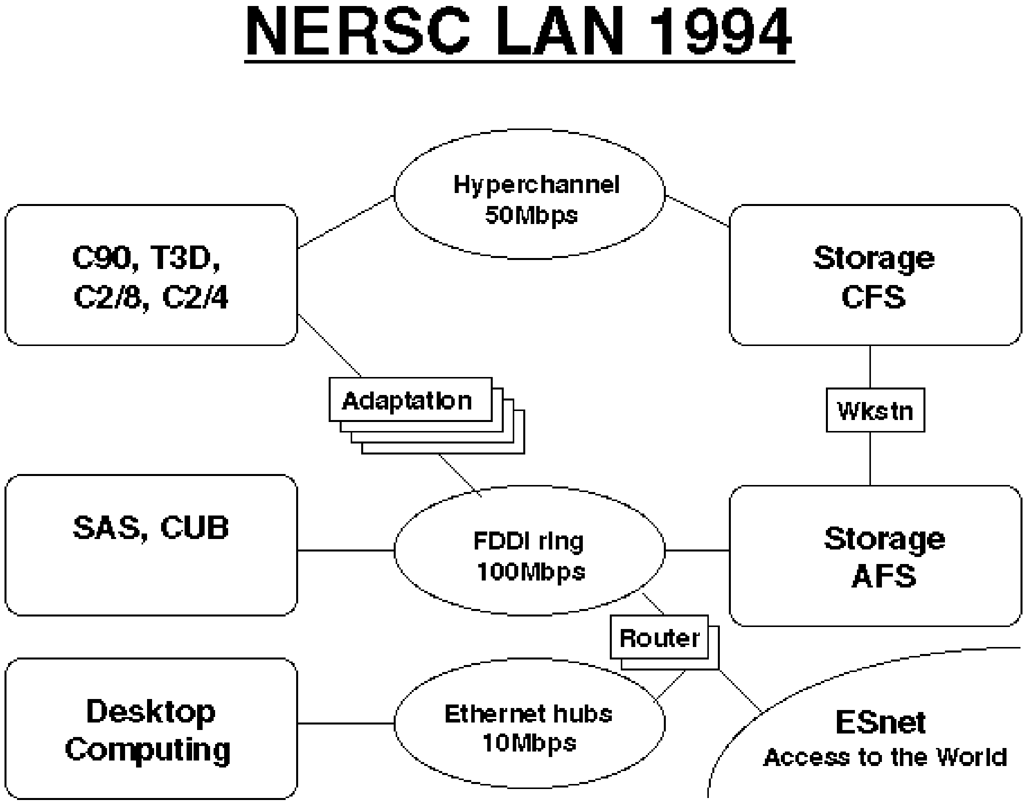


Fig. 2 - NERSC LAN mid-1994

# Next Phase of LAN Implementation

The local area networks at NERSC need to address and support certain broad goals:

***The Center should support an array of platforms, from the very large to the small and portable.***

***Continuous availability is an essential aspect of quality service.***

***A comprehensive Network Management System should be supported.***

***Traffic volume and pattern statistics must be available.***

***Switched or Shared media appropriate to the task should be supported.***

***Order of magnitude capacity insertion every 2 years is essential to keep up with latest technological developments including the pressure from MPP grand challenge scale applications.***

.

***Advanced deployment testing of next generation equipment is an important aspect of staying current and competent.***

## Milestones (By 1995)

As we move forward from this point, the implementation plan targets the high end to complete the basic architecture and bring it up to the necessary performance levels.

1. Top Tier: At the top tier interconnection of the large scale computing and file storage systems requires non-blocking media running at the speed of the connected interfaces. The individual interfaces operate at 100 megabytes-per-second (800 Mbps). The Block/File segment will be initially implemented with an 8 port HIPPI switch where each port runs at the full channel rate. The Packet/Interactive segment is pushing the state of technology enough that products to implement it at these speeds do not exist at the time of this writing. There are development efforts underway where the technology being pursued is the industry standard Asynchronous Transfer Mode (ATM) as it emerges in local area application. The likely implementation of this will provide OC-3 ports, each running at 622 Mbps.

2. Second Tier: At the second level the server class systems will be interconnected with each other and to the outside world with an FDDI (100 Mbps) ring network. The number of devices and data volume will require that this ring actually be implemented as a multi-port FDDI switch connecting several concentrators. An 8 port switch will provide an aggregate capacity of 800 Mbps for this tier.

3. Third Tier: This tier is tied to the evolution of the desktop machines. The network

needs to grow to keep up with the newer systems which are capable of demanding more of it. Currently, within the wing and trailer closets, wiring concentrators have been installed to provide thinwire Ethernet (10 Mbps) connections to the office. Where higher performance is required these concentrators need to be supplemented with Ethernet switches to provide traffic isolation between offices. Segments for these offices would then be independently capable of the full 10 Mbps, and restricted only by the aggregate traffic on the access between the switch and router. A higher performance alternative is being explored by advanced testing of local ATM. This solution would require waiting to install sonet OC-3c ATM to those offices placing the highest burden on the shared Ethernet. While this has a higher initial cost it would prolong the investment in current Ethernet hardware while skipping the FDDI generation.

4. Remote Access: Travel and home systems are supported via asynchronous dial modems and ISDN Ethernet bridges. The V.34 (28.8 kbps) standard modems are on the verge of release and the existing modems need to be upgraded to that level. The number of ISDN connections will also increase requiring a local pool of bridges or one of the new PRI capable hubs.

Figure 3 below shows the short term plan for FY 95 implementation.

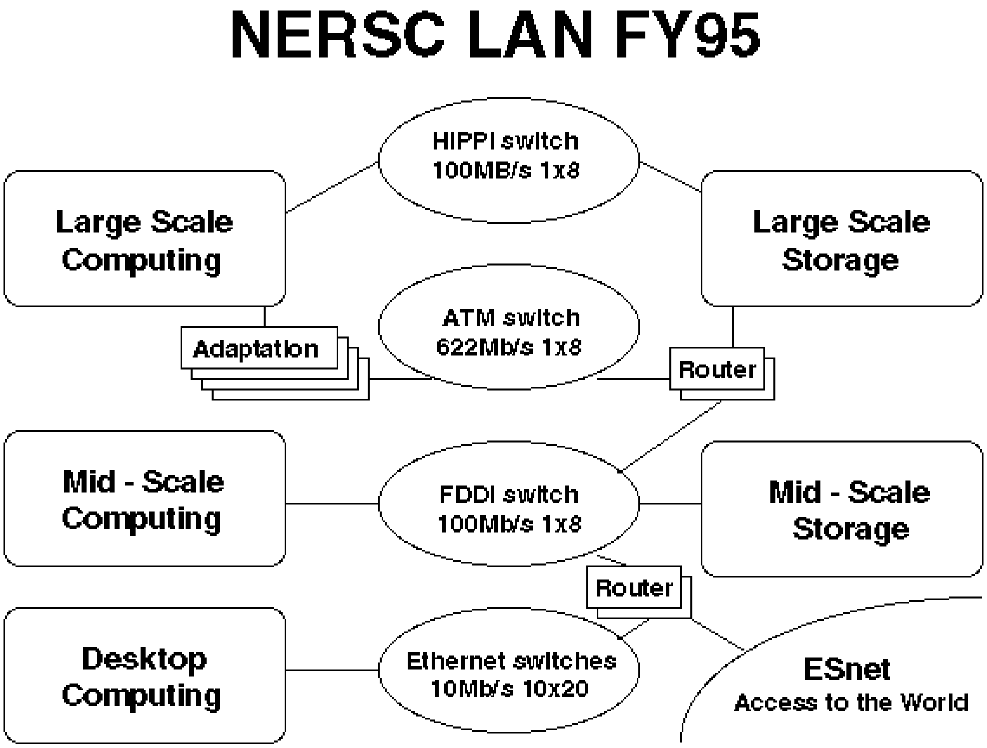


Fig. 3 - NERSC LAN Short Term Implementation

Tracking and Monitoring: (By early 1995) A Network Management System is required to track this assortment of equipment and provide the information necessary to operate a stable production service. By making this system a clone of that used to manage ESnet, the operator interface and system management overhead will lowered. Additional tools are necessary at each tier to provide technology specific diagnostic information. These tools include media level protocol analyzers, electrical or optical continuity and signal quality measurement devices, and test pattern generators.

Advanced deployment testing is being addressed via the recent LATM proposal [2] which seeks to position NERSC as a leader in the deployment of this state of the art technology. ATM has been selected for this effort because it represents a balanced approach to bandwidth and latency. The initial phase will focus at the workstation level with OC-3c (155M bps) interfaces. It will be connected back to the production LAN via a T3 (45 Mbps) interface on a third router attached to the FDDI ring. As the ATM to desktop technology becomes stable the advanced work will shift to OC-12c (622M bps) to fill in the top tier.

## Timeline

The ongoing nature of keeping the NERSC LAN up to date leaves the timeline in a state of continuous change. The Network Management Station will provide the necessary statistics to document the timing of additional needs. As such it needs to be deployed as soon as possible. Other high priority known bottlenecks are the Hyperchannel and the shared FDDI. These will be replaced in early FY'95.

## Summary

The first steps of the '91 plan were taken and resulted in a substantial gain in performance and productivity. However it was never completed and the high end was specifically left out. This has resulted in a bad situation getting worse over time with no end in sight.

Local computing and storage requirements continue to grow at tremendous rates. At the same time the wide area network capability is ramping up creating a state where even higher performance is required of the local NERSC networks. Given all this, the time is ripe for an aggressive approach to local area network development and deployment. The wide area bandwidth available to NERSC is already more than can be distributed locally, and is in need of this plan for resolving the situation.

[2] Proposal for NERSC Participation in an ESnet Community Advanced Technology Demonstration January 27, 1993

This document has presented updated planning and budget estimates to bring NERSC's LAN in line with the state of its other services. To survive as a premier computing facility, NERSC will need to address the limitations in the local network and it's continuing need for evolution.

# USER SERVICES

## Scope of Responsibility

We must provide effective user services which support the usability and accessibility of the high performance computing resources offered by NERSC, in a manner that is considerate of the needs and expectations of both NERSC and its customers.

## Area of Emphasis

The success of a centralized computing center like NERSC depends not only on the large scale computing resources it offers, but also on the supporting services it must offer to ensure that its intended users are aware of the resources available, that these resources are accessible, and can be easily and effectively used. These support services include consulting, documentation, and training. Adequate accounting and tracking must also be available to measure usage and provide justification to funding organizations. Two-way communication channels between NERSC and its customers must also be in place to provide customer support, encourage feedback, and publicize accomplishments. Flexibility in the methods used to access these services as well as the availability of a consistent and predictable interface to the services are also of great importance.

As NERSC evolves its computing resources to keep pace with rapidly changing computing technologies, our challenge is to also evolve and enhance the user services provided to ensure the continued usability and accessibility of the computing resources to its established and new user bases.

Several focus areas within NERSC services have been identified for this discussion:

- Customer Outreach
- Consulting
- Training
- User Interfaces
- Computer Allocation Accounting
- Information Preparation and Presentation Systems

Below we discuss the goals and milestones for each of these focus areas.

## Customer Outreach

To best serve our customers and meet their computing needs, NERSC must use a variety of methods to establish and maintain effective two-way communication between our customers and staff. Our methods must assure our customers that we are interested in and dedicated to meeting their unique computing needs. Through continuing visits to customer sites, including Town Meetings and customer attended conferences, we obtain first-hand knowledge of their work environment and needs, educate customers on methods that will help them to efficiently utilize the available computing resources through customer training, and maintain personal contact with our customers. As an organization we assimilate customer feedback, decide what actions to take, and provide feedback to the customers. We pursue existing customers to request their feedback on such topics as the quality and value of NERSC services, what software tools (public domain and commercial) are needed, and where these tools should be installed.

***Constantly seek new ways to improve the two-way communication between NERSC and the customers***

***Increase user participation in UNIXware project***

***Determine if NERSC should be trying to attract new customers and place more emphasis on marketing***

***Obtain more customer participation (reading and contributing) in bulletin boards, news, and the Buffer***

### Milestones

1. Town meeting reports: (By mid-1995) Make Town Meeting follow-up reports more widely available to everyone

2. Regional town meetings: (By mid-1995) Hold regional Town Meetings in an attempt to provide all customers with an opportunity to participate and to attempt to meet with smallergroups of customers

3. Staff rep: (By 1996) Appoint a staff representative to act as an agent for selected collaborations

4. Issue-focused conferences: (by mid-1995) Hold issue-focused conferences (e.g. a video teleconference on tools for parallelizing codes)

5. Wish list: (by mid-1995) Generate a "wish list" of requests collected via email, Town Meetings, etc., which can not be addressed immediately; regularly revisit the list.

6. User articles: (by 1996) Feature at least 3 articles written by users in the *Buffer* each year.

7. Make better use of ERDP data to develop an informational database about our customers

8. UNIXware tools (by 1996)
    - Include customer requests for purchased tools
    - Monitor software use to aid in future support decisions

- Determine if bulletin board is effective for polling, or if there are better ways
- Determine if we should wait for customer requests, or install on our own initiative

## Consulting

Participants in consulting at NERSC currently span many different groups.

The NERSC consultants are the on the front-line, answering phone calls and e-mail messages from our many customers. They are customer advocates, expeditors, and an information source for both NERSC customers and staff. As the most frequent interface to our customers, the consultants pass the feeling to our customers that NERSC is here to serve them and meet the computing needs of each customer in a timely manner.

The consultants provide a two-way interface between the staff and the customers, shielding the staff from time consuming interruptions to their daily work. The consultants take a proactive role in recognizing and anticipating problems before the customers experience the problem. Consultants track reported problems until they are resolved and then publicize the problem and its resolution to both NERSC staff and our customers. By maintaining a list of expertise at NERSC, the consultants are able, when necessary, to seek help from the appropriate people to answer a user question or request. The information is then funneled back through the consultants to the requestor. We make every attempt to *not* be a referral service, but to provide complete information and problem resolution services whenever possible. With each customer question or problem report, the NERSC consultants teach the customer how to find his/her own answers using the information sources and services that are available.

***Increase NERSC staff participation in consulting and resolving customer problems***

***Improve the information flow between the consultants and the NERSC staff so NERSC can be more proactive in problem detection and resolution, and in determining the customers needs***

### Milestones

1. <u>Expertise list</u>: (by 1996) Update and improve the list of expertise at NERSC to also reflect the services supported and the associated level of support

2. <u>Increase the number of staff participants</u> in our consulting services

3. <u>Expanded hours</u>: (by 1996) Evaluate whether to leave the consulting hours as they are (8-11:45am and 12:45-4:45pm, with the Operations staff available 24 hours/day for emergencies), or whether expanded consulting hours should be implemented

# Training

In addition to the training provided by the consultants during each contact with customers and through *Buffer* articles, the consultants and some other staff members present introductory classes. In some cases the classes are customized to match customer needs and, upon request, presented at a customer site. In other cases, classes are presented at NERSC by contractors and mainly attended by staff and customers from LLNL. Additional training sessions are presented at some Town Meetings when requested or offered when the need for them is observed by staff members.

***Use new technologies, such as video conferencing, to expand customer education possibilities (besides classes) and allow for different customer preferences and budgets***

***Conduct an annual survey to allow customers to select from list of possible future training classes***

***Provide details about other training opportunities for all of our customers***

**Milestones**

1. Pilot classes: (by 1996) Run pilot classes using other methods of presenting classes (i.e., video teleconferencing, two week institutes, visit customer sites, on demand tutorials possibly using Mosaic, demos on videotapes)

2. Explore alternatives to NERSC designing all classes
   - Possibly be a resource pointer to available outside training services and free training services available on the Internet
   - Organize more commercially available classes for our customers

3. Cookbooks: (by 1996) Develop cookbooks for graphics, applications, etc. to encourage self-training and provide new users with commands for basic functionalities

4. NERSC FAQs: (by mid-1995) Assemble our own FAQs (Frequently Asked Questions) for various computing topics

5. Internet FAQs: Point our customers to useful FAQs from the Internet by listing them in the *Buffer*, the NERSC home page, and bulletin boards

6. Train the Support group to handle more customer questions

# User Interfaces

This focus area encompasses all the approaches available to users when accessing

NERSC machines or services, and the visual and interactive aspects of these approaches. NERSC must develop a presentation of the NERSC services to our customers which is consistent and intuitive, and which enables NERSC to effectively implement the service usage policies.

***Provide one consistent, standard, predictable, and integrated X-windows approach to all tools, services, and platforms that NERSC provides.***

***Provide alternative approaches, if possible, to tools and services in addition to the selected standard approach.***

***Continue to provide motivation and education to our customers about the necessity of acquiring X-windows for their desktops.***

***Enforce through user interfaces any guidelines for the use of NERSC resources that we would like to see users follow, instead of expecting users to remember and abide by those guidelines. Included here are security, remote execution (license management), common home directories and load leveling.***

### Milestones

1. <u>Single service address</u>: (by 1996) Provide a single service address for all NERSC services. This includes one IP address for users to connect to (i.e through TELNET) for all NERSC services, and requires a single authentication and common home directory.

2. <u>Uniform style guide</u>: (by mid-1996) Establish a uniform style guide for NERSC service interfaces and enforce conformance for all NERSC supplied and supported services. This guide would also specify one standard NERSC graphical user interface.

3. <u>Standard application interface</u>: (by mid-1995) Develop a standard "wrapper" interface supported by NERSC to all vendor and NERSC supplied software/services. This wrapper will provide NERSC with, for examples, usage accounting, license management, and load balancing.

4. <u>Optional interfaces</u>: Where possible, provide other appropriate interfaces to services/ software. Options here include a command line interface or the "native" interface with which an application was developed.

5. <u>Foreign nationals</u>: (by mid-1995) Determine which services should be inaccessible to foreign nationals from sensitive countries. Enforce the guidelines developed for this purpose through user interface mechanisms in addition to the access right restrictions on files.

6. <u>X-windows support</u>: (by early 1995) Compile and expand on the information in the *Buffer* X-window series and make this available as a reference booklet for our staff

and users. Provide additional X-windows support using X-pert consulting, FAQs and help from other laboratory organizations like DCSP or Computations.

7. Dialup services: (by mid-1995) Decide if 800 dialup service via SLIP and X-remote will be offered to our users, and if so, how the issues of cost control and security will be addressed.

8. One-time passwords: (by mid-1995) Provide SecurID cards for people with special privileges and evaluate feasibility of providing them for all users.

## Computer Allocation Accounting

The Computer Allocation Accounting area covers the entire computer allocation cycle, from the solicitation and processing of proposals for allocations, through the implementation and enforcement of them, to the accounting of allocation usage. In all phases of this cycle, NERSC provides information in the form of reports to the appropriate people. In addition, NERSC supplies whatever utilities are needed to support the principal investigators, account managers and users throughout the allocation cycle.

Two major systems are supported by NERSC in order to meet the requirements of this area:

- The Centralized User Banking System (CUB) is the accounting system developed and maintained by NERSC to perform tracking and reporting of the allocation database. In addition to on demand reporting of user information, this system periodically generates accounting information, such as the monthly reports to PIs and DOE/OSC. Tools in the CUB system are also available to PIs and account managers for managing allocation disbursements to users in their account groups.

- The Energy Research Decision Package System (ERDP) is the means by which PIs generate and submit research proposals which require an allocation at NERSC. The system collects the proposals in a database from which reports are generated for review and evaluation by the supercomputing allocation committee at DOE headquarters.

***Add capabilities within the two systems above to automate parts of the allocation cycle that are still being handled manually by DOE staff and NERSC staff.***

***Provide means for PIs and account managers to perform functions that they should have privilege to perform instead of requiring that they channel these tasks through the NERSC Support staff.***

### Milestones

1. Foreign nationals: (by mid-1995) Implement the NERSC policy guidelines for

accounts for foreign nationals in both ERDP and CUB. In particular, add questions to ERDP regarding foreign national status and track the users in the accounting system.

2. Make the following functions available to the PIs: (by 1996)
   - add new users or request to add users
   - reset passwords for users in their repository group
   - clean up dead accounts, which includes an option to change the ownership of CFS files so that ownership of CFS files from dead accounts can be reassigned to another user if necessary.

3. Allocation history: (by 1996) Make information from *previous* allocation periods available on demand to, if not all users, at least the PIs by user or by repository.

4. Allow users to update all their own CUB information.

5. Make CUB user information tools accessible via WWW browser.

6. All year ERDP: (by mid-1995) Enable users to make year-round requests for allocations via ERDP and make the decision packages available online for SAC members all year as well.

7. Tool to set allocations: (by mid-1995) Add the capability of allowing DOE Program Managers to set allocations via an online tool. Currently, Linda Twenty, in Tom Kitchens office, collects all the authorized allocations from the DOE Program Managers, then enters them manually.

8. Terminate character interface support: (by mid-1996) To provide motivation for the PIs to acquire X-windows capability, at some time we will remove support for the character interface to ERDP.

9. Allocation requests: Provide new users with more information regarding Cray resources (time, space) needed to run codes in comparison to other machines so that they can make realistic estimates in applying for allocations in Decision Packages. Investigate whether providing benchmarks from various systems is feasible and useful.

10. CRU: Determine whether the CRU unit for allocation is still a reasonable unit to use, and if not, select a new unit for allocation amounts.

11. Multiple repositories: Enable users to belong to multiple repository groups without requiring multiple login names.

12. New user packet: (by end of FY95) Before the end of each fiscal year, review the information included in the new user packet for completeness and currency.

# Information Preparation and Presentation Systems

This area covers all external communication of information from NERSC. The main purpose of this communication is to keep our users informed and to provide public relations and marketing information to the general public.

Current vehicles for distributing information to NERSC users include:

- Online man pages
- DOCUMENT documentation presentation tool developed by NERSC
- Vendor-supplied documentation tools, such as DOCVIEW (CRI), LROM (HP)
- NEWS
- Wall messages (operator alerts)
- Online help tools accompanying vendor-supplied and NERSC-supplied tools
- E-mail
- U.S. mail - mass mailings to users
- NERSC's morning meetings (mainly to communication info to NERSC staff)

Current vehicles for distributing information, not only to NERSC users, but also external audiences include:

- *Buffer* newsletter
- Consultants
- Bulletin Boards
- World Wide Web/Mosaic
- Gopher
- Lab-wide sources (i.e. Newsline)
- Classes
- Brochures
- Videotapes

Currently, information is made available through a variety of presentation systems, and is often times accessible from only a subset of NERSC systems. In addition, the documentation preparation method and format is not necessarily the same for all the different systems. For example, FORMAT is used to mark up documents prepared for DOCUMENT, NROFF/TROFF for MAN pages, WWW/Mosaic expects HTML files and the vendor-supplied tools expect their own internal document formats.

The following goals and milestones will address these problems.

### ***Adopt SGML as the source format standard for NERSC documentation.***

DOE has mandated that the Standard Generalized Markup Language (SGML) be adopted as a standard format for documentation. By conforming to this mandate NERSC can easily import vendor documentation and share documentation with other sites.

***NERSC must rebuild its very successful, but outdated, system of document preparation, storage, and presentation around the SGML source format and take full advantage of the emerging information management tools as well.***

These tools will benefit the documentation effort but can also be applied to the more general area of information management in support of collaborations. By converting the existing documentation data base to SGML, employing new document preparation tools to add to this single source of information, and providing a variety of ways for users to easily find and view the information of interest, NERSC can regain its position as a leader in on-line documentation.

### Milestones

1. <u>MAN pages</u>: (by 1996) NERSC must recognize the high usage and importance of the MAN pages by
   - developing a document plan for updating MAN pages and a procedure for the documentarians to follow to install new versions of MAN pages.
   - establishing conditional filters and keys to automatically apply local modifications to new versions of vendor MAN pages.
   - making MAN pages available from a WWW browser
   - resolving performance problems in accessing MAN pages stored in AFS

2. <u>DTD</u>: (by 1996) Convert existing on-line documents to SGML using a NERSC Document Type Definition.

3. <u>Preparation</u>: (by mid-1995)Evaluate document *preparation* software for generating new SGML documentation, acquire the selected software, and train the editors in the use of this package. The new system should have the following desirable features for documentarians to prepare more sophisticated, multimedia documentation:
   - support for graphics and evaluation of layout of document
   - allow local site implementation notes
   - include mechanisms to restrict access to documents
   - provide for conditional display of information

4. <u>Database</u>: (by mid-1996) Build a new documentation *database* which has the capability to include distributed documents which, for example, are maintained by a vendor at a remote site.

5. <u>Presentation</u>: (by mid-1995) Evaluate information *presentation* software which enables a user to search on-line information (including MAN pages, *Buffer* articles, on-line documents) using exact search terms, approximate terms (fuzzy-matching), as well as synonyms. The presentation system should translate the selected source information "on the fly" to the format required by a select number of widely used browsers. The browsers should include WWW browsers and possibly a NERSC-developed X-windows interface.

6. Vendor documentation: (by mid-1995) Investigate how to incorporate existing proprietary vendor software into the new NERSC information system and restrict access to it. In particular, we should evaluate the CrayDocs product from CRI as a possible means of merging Cray documentation with our SGML source database.

7. Increase the readership and readability of the BUFFER by
   - improving the BUFFER cover to draw attention to its contents,
   - making the table of contents more enticing,
   - publishing more series of articles which could then be parked in the NERSC WWW home page for future reference,
   - including more articles about our users' work (e.g., SPP projects),
   - using color more often (possibly quarterly, or in centerfold articles),
   - generating better online publicity about the BUFFER, and
   - referring users to the BUFFER whenever possible